þÿ Workers of the World ? A Digital Approach to C
þÿInternational Scope of Belgian Socialist Newspa

Daems, Joke

2019-09

þÿDaems , J , D'haeninck , T , Hengchen , S , Zere , T & Verbruggen , C
þÿthe World ? A Digital Approach to Classify the International Scope of
þÿNewspapers, 1885 1940 ' , Journal of European Periodical Studies , vo
99-114 . https://doi.org/10.21825/jeps.v4i1.10187

http://hdl.handle.net/10138/305412
https://doi.org/10.21825/jeps.v4i1.10187

'Workers of the World'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940

Joke Daems, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen

Journal of European Periodical Studies, 4.1 (Summer 2019)

ISSN 2506-6587

The *Journal of European Periodical Studies* is hosted by Ghent University

Website: ojs.ugent.be/jeps

To cite this article: Joke Daems, Thomas D'haeninck, Simon Hengchen, Tecle Zere, and Christophe Verbruggen, "'Workers of the World"? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940', *Journal of European Periodical Studies*, 4.1 (Summer 2019), 99–114

‘Workers of the World’? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940

Joke Daems*, Thomas D’haeninck*, Simon Hengchen^, Tecle Zere°, and Christophe Verbrugge
*Ghent University, ^University of Helsinki, and °AMSAB Institute of Social History
joke.daems@ugent.be

ABSTRACT

Socialism has always been strongly related to internationalism, yet the attitude towards and expression of internationalism has likely changed throughout the years. Events such as the First World War, the post-war revival of institutionalized internationalism and the increasing geopolitical tensions during the interwar period are likely to impact the degree of internationalism found in socialism. In this paper, we use digital tools to search for expressions of ‘banal’ internationalism and cosmopolitanism in Belgian socialist discourse from 1885 until 1940 by text mining two socialist newspapers: the French journal *Le Peuple* and the Dutch *Vooruit*.[1] We use Named-Entity Recognition (NER) and a collocation analysis to study international and cosmopolitan sentiments expressed in these periodicals to better understand the impact of historical events. In addition, we highlight some of the difficulties encountered in collecting and processing the relevant data.

KEYWORDS

Socialism, internationalism, OCR, Named-Entity Recognition, collocation analysis

1 We would like to thank the Amsab Institute of Social History and the Royal Library of Belgium for providing us with the source material. We are also grateful for Belgian Science Policy (BELSPO) grant n° BR/121/A3/TIC-BELGIUM.

Introduction

Socialism and internationalism are almost by definition intrinsically related concepts. The slogan 'Workers of the world, unite', coined by Marx and Engels in the *Communist Manifesto*, has been endlessly repeated within the socialist movements across the world. Many local and national International Workingmen's Association party meetings end their meetings with the anthem, *The Internationale*; containing key themes of workingmen's solidarity which extend to the entire 'human race'. Thus, workers' internationalism was not only an idea and a way of mobilizing and organizing, it was also a practice that could be found in daily activity, conference participation, sports events, etc. Through mutualism and early-modern notions of brotherhood, socialist or workers' internationalism are related, but at the same time also different from the much older ideological construction of cosmopolitanism. The concept of 'cosmopolitanism' cannot be reduced to a single conceptualization.[2] Steven Vertovec and Robin Cohen therefore offer several 'perspectives on cosmopolitanism' which cover its definition as a 'socio-cultural condition' or as a practice, but also encompass political views or projects.[3] Rather than dealing with predefined 'normative' varieties of cosmopolitanism, this article focuses on forms that can be described as narrative constructs.[4] In previous research, we identified two major belle époque and interwar varieties of the age-old cosmopolitan views that all share the idea that human beings are or should be citizens of a single world community: the openness to transnational cultural exchange, and the embrace of diversity. Attitudes towards internationalism and cosmopolitanism have changed over time due to changing geopolitical situations and organizational landscapes.[5] In the last quarter of the nineteenth century, cosmopolitanism also became increasingly juxtaposed with traditional national values, and increasingly associated with bourgeois metropolitan culture and urban modernity.

In line with the rise of social democracy and a growing self-awareness in the last quarter of the nineteenth century, we might thus expect various degrees of openness and changing conceptions of internationalism and cosmopolitanism over time within the public discourse of the socialist movement. Moreover, internal rifts within socialist internationalism resulted in multiple social democrat, anarchist, communist, and even non-western socio-cultural constructs of internationalism.[6]

Newspapers were crucial for the creation of imagined national communities.[7] Socialist communities used newspapers to communicate to their public as a method of 'banal nationalism',[8] referring to the everyday representations of the nation: sporting events, national hymns, even a weather forecast (e.g. 'typical Belgian weather'), which imply togetherness. This serves to create a division between 'international' and 'domestic' news.

2 Steven Vertovec and Robin Cohen, 'Introduction: Conceiving Cosmopolitanism', in *Conceiving Cosmopolitanism: Theory, Context, and Practice*, ed. by Steven Vertovec and Robin Cohen (Oxford: Oxford University Press, 2002), p. 5.

3 Vertovec and Cohen, pp. 9–15. See, for instance, Garrett Wallace Brown and David Held, *The Cosmopolitanism Reader* (Cambridge: Polity, 2010).

4 Maria Rovisco and Magdalena Nowicka, eds, *The Ashgate Research Companion to Cosmopolitanism* (Farnham: Asghate, 2011).

5 Elaborated in Daniel Laqua and Christophe Verbruggen, 'Beyond the Metropolis: French and Belgian Symbolists Between the Region and the Republic of Letters', *Comparative Critical Studies* 10.2 (2013): pp. 241–58. In his discussion of a globalized world, Ulrich Beck promotes a 'cosmopolitan vision' that acknowledges diversity and cultural exchange: *Der kosmopolitische Blick oder: Krieg ist Frieden* (Frankfurt am Main: Suhrkamp, 2004).

6 Patrizia Dogliani, 'European Municipalism in the First Half of the Twentieth Century: The Socialist Network', *Contemporary European History,* 11.4 (2002), 573–96.

7 Timothy Baycroft, *Culture, Identity and Nationalism: French Flanders in the Nineteenth and Twentieth Centuries* (Suffolk: Boydell & Brewer, 2004), pp. 148–67.

8 Michael Billig, *Banal Nationalism* (London: Sage, 1995).

Both the shift in orientation and the use of the press in communicating with sympathizers is apparent in the case of Belgium in this period. Contrary to the myth that Belgian socialists have always been oriented towards the Belgian 'fatherland' in which they could realize their social democratic aspirations, already before 1914 a strong regionalist tendency existed among Flemish and French-speaking socialists, mingled with an 'internationalist' and Belgian orientation.[9] Less studied is how the First World War, the post-war revival of institutionalized internationalism, and the increasing geopolitical tensions during the interwar period in Belgium influenced this shift. Do we notice time periods in which Belgian socialists were exposed to internationalist thinking through subtle expressions of cross-boundary solidarity and class consciousness?

We are thus interested in expressions of 'banal' socialist internationalism, in comparison with national sentiments, that can be found in socialist newspapers in Belgium in the period 1886–1936. We consider two socialist daily newspapers in the French and Dutch speaking regions: *Le Peuple* and *Vooruit*, respectively. Without high quality structural metadata, it is hard to measure the relative weight and presentation of 'international' or domestic news, therefore it is necessary to explore other methodological pathways. Using a distant reading approach of a digital corpus of OCRed (optical character recognition) copies of the newspaper, we combine frequency analysis of the use of banal and international concepts in texts, and a historical mapping of the terms over time to study the abovementioned long-term changes. In documenting this approach, we offer both insights into the international scope of Belgian socialist newspapers, and the methodological considerations. This introduction is followed first by a methodology section, where we introduce the data collection and cleaning, including a discussion of the OCR issues, followed by a discussion of the two methods that we applied: identifying the local, national, and international news items through the use of Named-Entity Recognition (NER), i.e., identifying and extracting geographical entities, and assessing the internationalist and cosmopolitan rhetoric in the newspapers' discourse through word collocation. After the methodology, we describe the most important results from these analyses, and reflect on our findings in the discussion and conclusions section. This article furthers the work of Van de Bos and Giffard, who have been studying emerging nationalism in public discourse,[10] and generates insight into the use of digital periodicals as a medium for historical research.

Method

Corpus creation and OCR quality checks

We selected nine samples of digital editions of two major socialist periodicals in the nineteenth century, covering time periods that have been crucial in the socio-political history of Belgium and Europe as a whole: 1886–87, 1892–93, 1901–02, 1905–06, 1912–13, 1917–18, 1924–25, 1931–32, and 1937–38.[11] The first, *Le Peuple*, was a socialist daily newspaper published in Brussels, Belgium, starting in 1885. For a long time, it

9 Maarten Van Ginderachter, 'Social Democracy and National Identity: The Ethnic Rift in the Belgian Workers' Party (1885–1914)', *International Review of Social History,* 52.2 (2007), 215–40; Maarten Van Ginderachter, 'Contesting National Symbols: Belgian Belle Époque Socialism Between Rejection and Appropriation', *Social History,* 34.1 (2009), 55–73.

10 Maarten van den Bos and Hermione Giffard, 'Mining Public Discourse for Emerging Dutch Nationalism', *Digital Humanities Quarterly,* 10.3 (2016), 87–99.

11 The newspaper *Vooruit* was recently digitized by two Belgian institutions, namely Amsab Institute of Social History and the Royal Library of Belgium (KBR). *Le Peuple* was also obtained courtesy of the Royal Library of Belgium. However, there were gaps in the data: for *Vooruit*, no data were available for the years 1905 and 1918; for *Le Peuple* data were missing for the years 1917 and 1918 (except for months 11 and 12 of 2018).

was considered the central organ of the Belgian Labour Party. The second, the Ghent based *Vooruit*, started one year earlier, in 1884. The socialism in Ghent was characterized by a local neighbourhood sociability and a strong organizational affection for the local consumer cooperative, also called *Vooruit*. The socialist movement in Brussels was less proletarian and historically firmly rooted in a cosmopolitan and intellectual liberal tradition, which makes the comparison even more interesting. However, the successful Ghent cooperative movement led early Belgian socialism and the Ghent model resonated well beyond the Belgian borders. Both movements in Ghent and Brussels have always shown a strong affinity for international social democracy, notably the German workers' party. It is nevertheless impossible to discuss notions of internationalism without taking into account expressions of nationalism. Access to the samples was provided by the Royal Library of Belgium (KBR) and the AMSAB Institute of Social History, for which we are grateful, still, it took a few months to collect the required legal permissions and signatures to get this access in the first place.

After the data acquisition, other methodological concerns cropped up, caused by the OCR quality and the lack of metadata. A preliminary assessment was carried out on the OCR quality and the need for re-OCR. Samples consisting of single newspaper pages were randomly selected for each of the newspapers *Vooruit* and *Le Peuple*. The OCR layer on these pages was manually corrected to serve as ground-truth material. The images for the same newspaper pages (tiff for *Vooruit* and PDF for *Le Peuple* as we did not yet have tiff files) were then uploaded in Transkribus for re-OCRing using the available Transkribus OCR models. The models include that of Abbyy OCR engine (Abbyy model) and a custom model developed by the National Library of France (BnF model). The Abbyy model is a standard OCR model used in Transkribus, whereas the BnF model is a custom OCR model used for French-language newspapers. The OCR quality of the existing and newly generated text was compared with the ground-truth text using the Levenshtein distance, which is a measure of the similarity between two strings of text.[12] The result of this test (Table 1) was an improvement of less than 2%, which made us conclude that the effort of improving the OCR of the whole corpus would be very unrewarding. Therefore, we continued to work with the original OCR, being aware of the differences between *Le Peuple* and *Vooruit* in OCR quality.

Table 1 An assessment of the accuracy of existing (KBR-OCR and Amsab-OCR), and newly generated OCR by Transkribus (Abbyy-reOCR and BnF-reOCR), for two Belgian historical newspapers *Le Peuple* (French) and *Vooruit* (Dutch)

***Le Peuple* (10557 characters)**			***Vooruit* (9797 characters)**		
	Levenshtein	*%*		*Levenshtein*	*%*
KBR-OCR	2480	**76.5%**	*Amsab-OCR*	580	**94.1%**
Abbyy-reOCR	2763	**73.8%**	*Abbyy-reOCR*	412	**95.8%**
BnF-reOCR	2380	**77.5%**	-	-	-

The documents scanned by Amsab Institute of Social History seemed to be of higher quality than those of *Le Peuple*. In the case of the *Le Peuple* corpus, it is likely that the bad OCR quality could only be improved by rescanning the sources in line with current standards and technologies, which was beyond the scope of the present study. Part of the explanation is the difference between French and Dutch and the fact

12 Michael Gilleland, 'Levenshtein Distance, in Three Flavors', *Merriam Park Software* (2009).

that French has more diacritics (é à ç è ù î ë). This makes things harder for the OCR engine, especially in printed newspapers where the font size is small. These factors could affect the NER results, although it has been indicated by Rodriquez et al. that manual correction of OCR output does not significantly improve the statistical relevance of NER performance.[13] On a broader note, a similar conclusion about the impact of OCR on different historical text analyses has recently been carried out. Mark J. Hill and Simon Hengchen argue that for some analysis the OCR quality is perhaps less problematic than one may initially guess.[14] Nonetheless, OCR quality depends on various factors (such as font, quality of the scan, quality of the paper, smudges, damage, or shine-through), and the two studies above have worked with different kinds of material (respectively mid-twentieth-century typewritten transcripts and eighteenth-century books), which could explain why our experience does not match theirs.

As Dan Edelstein has argued, the digital age has been a boon for intellectual historians who can, due to the mass digitization of historical texts, do their research more efficiently than ever before.[15] The information they usually need revolves around concepts, in our case expressions of 'banal' socialist internationalism and national sentiments, and larger information chunks or relationships that involve concepts. From a distant reading perspective, a particular concept consists of tokens (words, phrases, or even whole sentences) that occur in texts and refer to this concept. The difficulty one thus needs to deal with is searching for ways to retrieve all the multiple ways a given concept may be referred to in a text, such as synonyms, spelling variations, abbreviations, multilingualism.[16] In addition to this, one needs to be aware of the time-sensitiveness of historical texts and the semantic changes that may occur. Also, concepts can be controversial and therefore contested via rival definitions of the same word in public discourse over time.[17] In reality, all the errors that are produced during the OCR process are an additional difficulty one needs to deal with and can overcome but are arguably the hardest one.[18]

Identifying national and international news items via Named-Entity Recognition analysis

To study to what extent socialist periodicals published news items related to local, national, or international events, we identified the different cities, countries and regions mentioned in the newspapers through Named-Entity Recognition (NER). NER is a way of extracting and classifying named entities such as the name of persons, organizations, expressions of time or, in this case, locations. Prior to the NER, the texts needed to be

13 Kepa Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska, 'Comparison of Named Entity Recognition Tools for Raw OCR Text', *Proceedings of KONVENS* (2012), pp. 410–14.

14 Mark J. Hill and Simon Hengchen, 'Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study', *Digital Scholarship in the Humanities*, in production, fqz024.

15 Dan Edelstein, 'Intellectual History and Digital Humanities', *Modern Intellectual History,* 13.1 (2016), 237–46.

16 Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou, 'Text Mining the History of Medicine', *PloS ONE,* 11.1 (2016): e0144717.

17 Edelstein, p. 239. See also Peter De Bolla, *The Architecture of Concepts: The Historical Formation of Human Rights* (New York, NJ: Fordham University Press, 2013); and Reinhart Koselleck, 'A Response to Comments on the Geschichtliche Grundbegriffe', in *The Meaning of Historical Terms and Concepts*, ed. by Hartmut Lehmann and Melvin Richter (Washington DC: German Historical Institute, 1996), pp. 59–70.

18 Matthew Jockers and Ted Underwood, 'Text-Mining the Humanities', in *A New Companion to Digital Humanities,* ed. by Susan Schreibman, Ray Siemens, and John Unsworth (Hoboken, NY: Wiley-Blackwell, 2016), pp. 291–306.

cleaned and filtered for processing. We processed the PDF scans and XML ALTO files.[19] While there are scripts and tools available to do this, many of them do not work well with larger documents, let alone on individual computers. Ideally, such processes are run on servers with greater capacity, which researchers do not always have access to, or they lack the knowledge needed to use them effectively. Even if they do, running the scripts and processes can still take more than a few days, depending on the size of the corpus, the efficiency of the scripts and the computational power of the machines running them. For the actual extraction and disambiguation of locations (geographical entities) the software package MERCKX (Multilingual Entity/Resource Combiner & Knowledge eXtractor) was used. MERCKX was recently developed 'to extract entity mentions from documents and to link them to DBpedia'. This workflow consists of downloading resources, building the dictionary, tokenizing the text, and spotting location mentions.[20] This solution was chosen for four reasons. First, to the best of our knowledge, this is the only tool that has been tested on (and, actually, developed for) historical newspapers in Dutch and French where entities can be referred to in different languages within the same newspapers. Second, due to its reliance on an external knowledge base, it does not require training and can take into account historical spellings (such as *Passendale* vs *Passchendaele*, two spellings for the same Belgian village). Third, the tool is shown to perform better than other, state-of-the-art NER tools on cultural heritage data. Finally, MERCKX is available as open source software, allowing us to customize it to our needs, which is what we did: the python script which extracts all URIs matching a given type in DBpedia was slightly modified to apply for multiple files and was run to extract the DBpedia type of location (LOC).

During the NER processing, we discovered a few issues. The city of Brussels, for example, scored very low, supposedly barely appearing in either *Vooruit* or *Le Peuple*, although it is the capital of Belgium and likely to appear very frequently in both newspapers. This was caused by the way the data is organized in DBpedia. That is, the DBpedia resource does not list *Bruxelles* and *Brussel* as possible labels. Instead, it lists *Région de Bruxelles-Capitale (fr)* and *Brussels Hoofdstedelijk Gewest (nl)* which almost never appear in our newspapers. The results also showed the importance of removing stop words before the analysis of toponyms, as a number of stop words were recognized as toponyms, for example, *Aan* as a river in New Zealand, and *Als* as an Island in Denmark, presumably because they appeared at the start of a sentence and thus had a capital letter.

With the improved NER processing carried out and verified, the extracted named locations were then grouped per year and per sample period for further analyses. Furthermore, a Python script was written to extract the corresponding country and continent names for all the extracted named locations. An overview of the coverage for *Vooruit* and *Le Peuple* can be seen in Tables 2 and 3 below. In these tables, the number of characters and words in the input text, and the number (and percentage) of identified locations has been presented per sample period. As is shown in the tables, on average more than 1% of the input number of words were identified as locations (1.3 % for *Vooruit*, 1.2% for *Le Peuple*).

As we had discovered to be necessary in the first NER run, we removed common words that were labelled as toponyms as an additional cleaning step (for *Vooruit*: *Bij, Noord, Noorden, Juist, Belg, Waarde, Zou, Tegen, Weer, Binnen, Heeren, Vereeniging, Hét,*

19 For the sample years, a total of 29,875 XML files were obtained for *Le Peuple* and 34,597 XML files for *Vooruit*.

20 Max De Wilde and Simon Hengchen, 'Semantic Enrichment of a Multilingual Archive with Linked open Data', *Digital Humanities Quarterly*, 11.4 (2017).

Table 2 Overview of number of words, characters, and locations in the dataset per sample period in *Vooruit*

	Input data		Output data (locations)	
Sample period	**Number of characters**	**Number of words**	**Number of locations**	**% locations**
1886–87	25,820,938	4,125,201	33,538	**0.8%**
1892–93	36,966,563	6,420,170	39,462	**0.6%**
1901–02	63,422,683	10,900,773	84,109	**0.8%**
1906	48,748,611	8,489,781	62,355	**0.7%**
1912–13	103,601,481	16,516,139	158,549	**1.0%**
1917	20,072,467	3,151,705	38,732	**1.2%**
1924–25	74,492,144	11,813,512	154,103	**1.3%**
1931–32	155,714,651	24,904,874	363,683	**1.5%**
1937–38	193,268,465	30,648,144	469,503	**1.5%**
Total	**659,320,502**	**106,424,928**	**1,331,034**	**1.3%**

Table 3 Overview of number of words, characters, and locations in the dataset per sample period in *Le Peuple*

	Input data		Output data (locations)	
Sample period	**Number of characters**	**Number of words**	**Number of locations**	**%locations**
1886–87	24,235,445	4,749,005	41,093	**0.9%**
1892–93	58,719,597	11,670,992	148,256	**1.3%**
1901–02	78,619,632	15,366,417	174,548	**1.1%**
1905–06	67,467,419	13,229,181	129,397	**1.0%**
1912–13	102,988,739	20,661,437	182,394	**0.9%**
1918	2,759,493	544,970	4,829	**0.9%**
1924–25	110,593,463	22,292,955	252,220	**1.1%**
1931–32	131,970,111	27,144,864	321,923	**1.2%**
1937–38	155,024,415	31,599,229	474,239	**1.5%**
Total	**649,423,272**	**130,839,053**	**1,539,550**	**1.2%**

Welk, Zeven, Samen; for *Le Peuple*: *Afin, Aucun, Belg, Voies*).[21] An additional column was added to the data to assign the broader region (Belgium/Neighbouring country/Europe/continent) to the countries that made up more than 0.5% of all toponyms.

Grasping international and cosmopolitan sentiments via collocation analysis

Several distant reading techniques can be used to grasp the changing sentiments towards socialist internationalism. One such technique is topic modelling, a text mining method that uses statistics for the discovery of hidden semantic structures in a corpus by

21 'Belg' arguably expresses a nationalist sentiment, but it is not a place name and as such will not be taken into account for the toponym analysis. These expressions were included in the collocation analysis in part 2 of the paper.

identifying groups of words that occur relatively frequently together.[22] While often used in digital humanities, topic modelling works best for finding hidden topics that are not predefined by a researcher.[23] In our case, however, we needed predefined topics to study the occurrence of banal internationalism. Rather than using topic models, our approach consisted of three steps. In the first, we defined our own topics of nationalism and internationalism with key concepts that we expected to be associated with internationalism: obviously the concept of internationalism itself, but also proxy-concepts such as universalism and cosmopolitanism.[24] We looked at how frequently these concepts occurred in both the corpora of *Vooruit* and *Le Peuple*. In the second step we used Antconc, a freeware corpus analysis toolkit for concordancing and text analysis, to query for collocations with our predefined topics, within a range of five tokens left and right. This statistical analysis searches for words that are most likely to occur in the close proximity of the topics internationalism and nationalism, compared to its occurrence elsewhere in the corpus. This results in a list of statistically significant associations for each of the sample years. Next, this list was (partially) manually cleaned. Variations in spelling, OCR-mistakes, plurals, and declensions were normalized to the radix (root word) or the token that occurs most frequently in the corpus. In the third step, we selected the top 100 results for each sample year and combined them in two groups, before and after the First World War. Next, we uploaded the combined results into Wordart to generate a word cloud to create a better overview of those most important words associated with the pre-selected topics. Ultimately, this allowed us to compare *Vooruit* with *Le Peuple* in terms of associations with internationalism and nationalism.

As a final step of our attempt to grasp banal socialist internationalism, we performed an analysis of pronouns and expressions such as 'our French brothers' and 'our foreign friends'. Though we assumed that these would reveal the existence of a banal socialist internationalism and sense of solidarity in *Le Peuple*, our query search resulted in hardly a handful of hits. Expressions such as *ons vaderland* (our fatherland) or *ons Vlaanderen* (our Flanders) also yielded hardly any results in *Vooruit*, and we know from secondary literature that they should be included in our sample. Here, perhaps, we should question the quality of the OCR data again. Clearly, it is harder to retrieve phrases instead of words when running collocation queries, presumably because there is a higher chance of errors in the OCR output. As has been argued before, we can confirm that bad OCR is more problematic for a collocation analysis than for NER. OCR corruption leads to both losing interesting collocations, and generating insignificant collocations of non-existing words through the introduction of noise.[25]

22 This results in a list of keywords that, together, represent a topic. These keywords are supposedly the most representative tokens for that topic: with them combined, a human operator must deduce the underlying theme. Existing applications show that this is often more of an art than a science. See, for example, Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei, 'Reading Tea Leaves: How Humans Interpret Topic Models', *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (2009), pp. 288–96.

23 Robert Nelson, *Mining the Dispatch* [accessed 5 November 2018]; David Newman and Sharon Block, 'Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper', *Journal of the American Society for Information Science and Technology,* 57.6 (2006), 753–67; and Mathew Jockers and David Mimno, 'Significant Themes in 19th-Century Literature', *Poetics,* 41. 6 (2012), 750–69. For a broader overview of topic modelling within historical research, see Rene Brauer and Matz Fridlund, 'Historicizing Topic Models: A Distant Reading of Topic Modeling Texts within Historical Studies', *International Conference on Cultural Research in the Context of 'Digital Humanities'* (St. Petersburg: Russian State Herzen University, 2013).

24 Exact terms of the topics: list of keywords **nationalism** (Dutch): vlaande* vlaams* ; Waals*; Wallo*; belgi* nationalism*; vad*rland*; list of keywords **internationalism** (Dutch): *osmop*; internatio*; buitenla*, overzeesch*n; univers*; list of keywords **nationalism** (French): belg*; nationa*; patri*; wallo*; flaman* list of keywords **internationalism** (French): univers* *osmop* international* *trang*.

25 Hill and Hengchen.

Results

NER analysis

We looked at the 100 most common toponyms in *Vooruit* and those in *Le Peuple* and studied their frequency throughout the years. *Vooruit* seemed to really be focused on Ghent (14.36% on average), followed by Belgium, Paris, and France with much lower percentages (3.38%, 3.32%, and 3.11%, respectively). Striking changes in frequency can be found for Ghent, accounting for 24.82% of all toponyms in 1901 and 1902 but only for 7.95% in 1937 and 1938, Berlin (on average accounting for 1.13% of all mentions, but increasing to 4.85% in 1917), Russia (0.98% on average, 3.71% in 1918), and Ostend (on average 2.05%, going up to 4.28% in 1901–02 and down to 0.15% in 1917).

In contrast to *Vooruit*, where our analysis indicated a clear local focus of the newspaper, *Le Peuple* had an international outlook. For *Le Peuple*, the focus is first on Paris (5.1% of all toponyms), but the difference between the most frequently mentioned place and the other frequently mentioned places is not as great as it is in *Vooruit* (where Ghent makes up 14.36% of all places, and the next most common place makes up 3.38%). For example, the second most common place name is Belgium (4.57%), followed by France (4.02%), and Antwerp (3.76%). The frequency throughout the years seems mostly comparable as well. Notable exceptions are 1901–02 and 1905–06, where Russia gains presence (2.53% and 2.49% of all mentions compared to the overall 1%), and 1918, where the impact of the First World War is clearly noticeable. Here, mentions of Belgium rise to 10.11% (compared to the overall 4.57%), followed by Germany (7.3% compared to the overall 1.85%) and Berlin (4.49% compared to the overall 1.3%), while other places become less frequent than in other years.

We also studied the top twenty countries overall and compared those across the years. The changes in the top eight after Belgium can be seen in Figs 1 and 2. Both *Vooruit* and *Le Peuple* focus most heavily on Belgium and France (62% of all toponyms are from these countries), but the focus on Belgium is greater in *Vooruit* than in *Le Peuple* (48.31% compared to 39.95% of all mentions). The top five countries in *Vooruit* is further completed by the United Kingdom (6.14%), the Netherlands (5.62%), and Germany (4.92%). *Le Peuple* shares Germany (5.17%) and the United Kingdom (4.46%) but contains more mentions of Italy (3.56%) and even Spain (2.7%) than the Netherlands (1.96%). This can in part be explained by the fact that The Netherlands is a Dutch-speaking country (and thus more important to a Dutch-speaking audience), although it is still a neighbouring country of Belgium.

The most striking difference for *Vooruit* is to be found in the year 1917, when less than 30% of all toponyms are from Belgium, and there is an increased number of references to United Kingdom (9.05% compared to the overall 6.14%), Germany (8.6% compared to the overall 4.92%), Russia (4.04% compared to the overall 1.54%), South Africa (3.42% compared to the overall 1.32%), and Sweden (4.11% compared to the overall 0.75%). For *Le Peuple*, Belgium received less attention in 1901–02, 1905–06, and 1918. In 1901–02 and 1905–06, more attention goes to Italy (up to 5.43% and 4.01% compared to the overall 3.56%) and Russia (up to 3.17% and 4.69% compared to the overall 1.67%), which can be related to the political turmoil. 1901–02 also sees an increased importance of Hungary and Portugal, whereas 1905–06 sees an increased importance of Japan during and in the aftermath of the Russo-Japanese War. In 1918, the most striking rise is in the expected number of mentions from Germany (15.77% compared to 5.17% overall), while in 1937–38 the Spanish Civil War received a lot of attention among Belgian socialists (5.48% compared to 2.7% overall).

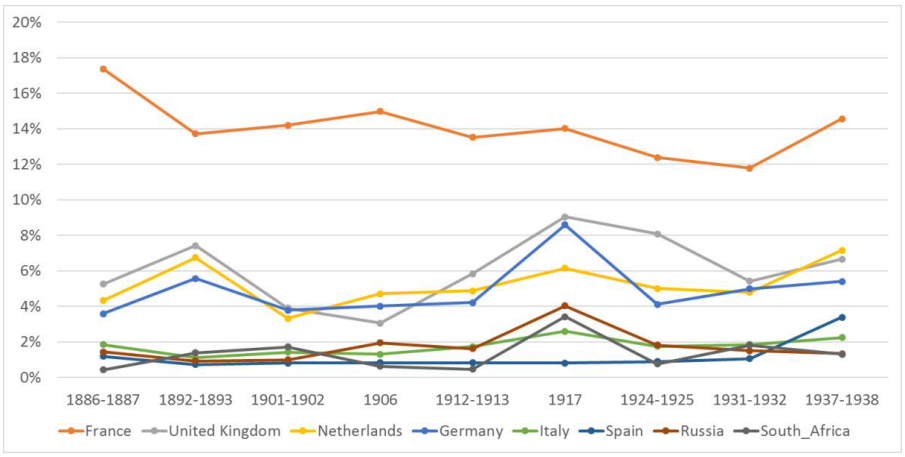

Fig. 1 Most frequent countries of origin in *Vooruit* throughout the years (excluding Belgium, 1886–1938)

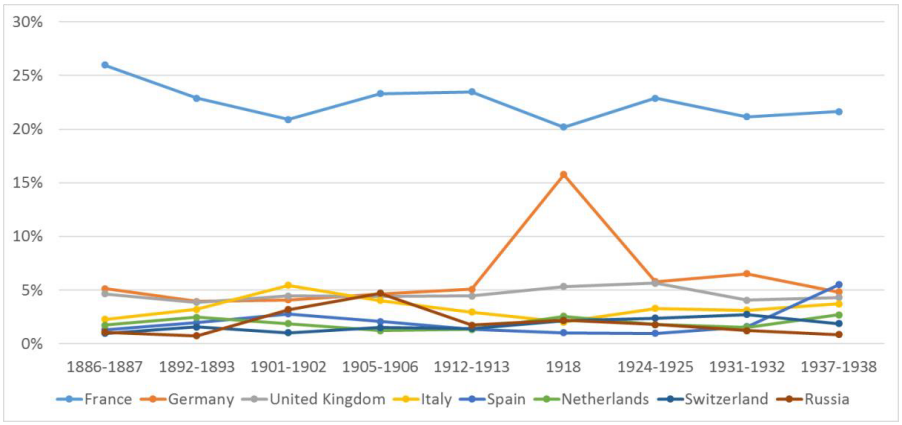

Fig. 2 Most frequent countries of origin in *Le Peuple* throughout the years (excluding Belgium, 1886–1938)

When looking at the greater regions in Fig. 3, *Vooruit* generally focuses more on Belgium than on any other region (50.1% overall), except for 1917. Neighbouring countries receive the most attention (31.32% overall), followed by Europe (13.19%). There is somewhat more attention to neighbouring countries in 1892–93 and 1937–38 and more attention for neighbouring countries and Europe in the War Period 1917–18. In 1901–02 and 06, mentions of Belgium make up around 60% of all mentions, an increase of 10% compared to the overall mentions.

Le Peuple tells a somewhat different story. As can be seen in Fig. 4, here as well, the focus is on Belgium, but the difference between Belgium and the neighbouring countries is not as great as for *Vooruit* (40.95%–34.61% vs. 50.11%–31.32%). In 1917–18, there are more mentions of neighbouring countries than there are of Belgian places. Where in *Vooruit* this peak was mimicked by a rise in mention of European place names, this trend cannot be found in *Le Peuple.*

Collocation analysis

We first looked at the frequency of key concepts related to nationalism and internationalism in our corpus. As can be seen in Fig. 5, concepts related to nationalism occur more frequently than those related to internationalism for both newspapers.

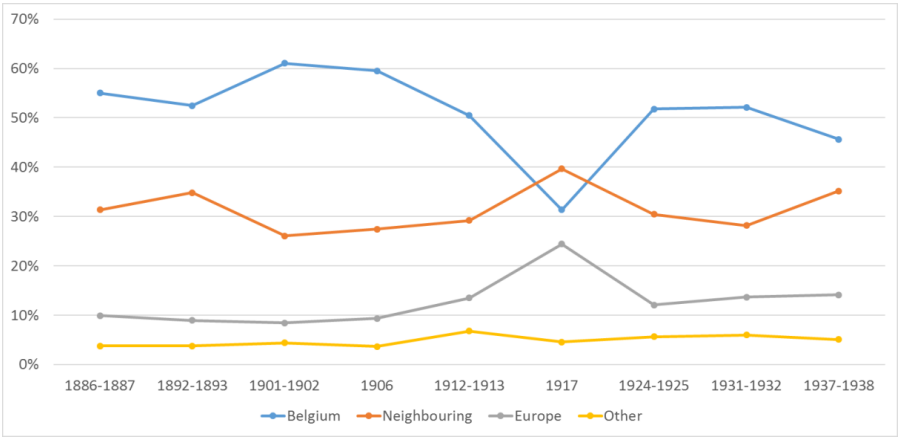

Fig. 3 *Vooruit* — focus on Belgium versus other regions (1886–1938)

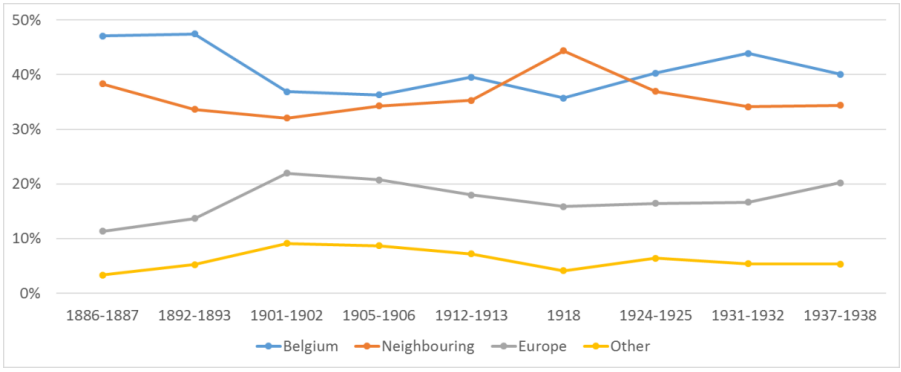

Fig. 4 *Le Peuple* — Focus on Belgium versus other regions (1886–1938)

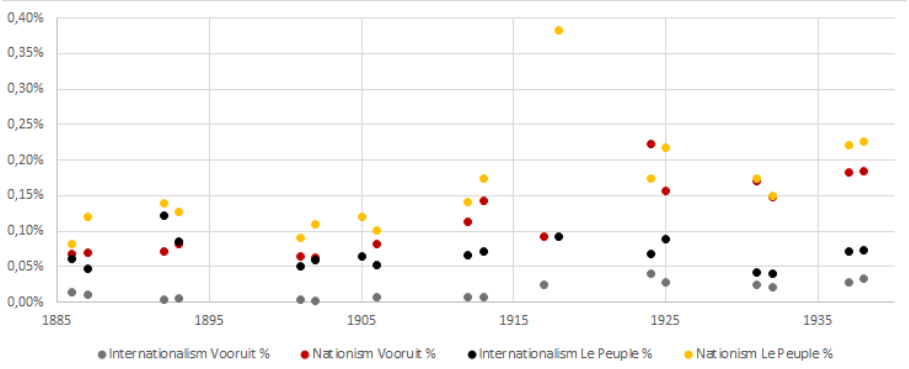

Fig. 5 Comparison occurrence topic nationalism and internationalism in *Le Peuple* and *Vooruit* (1886–1938).

There is a steadily growing focus on nationalism with peaks in the interwar period. Especially striking are the years 1917 and 1918, where in *Vooruit* the distance between nationalism and internationalism is much smaller than in the surrounding sample years, mimicking the change in focus on Belgium versus other regions depicted in Fig. 3 in the NER analysis for that year, but particularly *Le Peuple* clearly shows much stronger nationalist sentiment in 1918.

In the next step, we used the lists of 100 words that were most strongly associated with our topics of nationalism and internationalism according to the statistical analysis with Antconc, to generate word clouds depicting the most common word associations in the period before and after the First World War. In general, as can be seen in Figs

6–9, most of the associations with the topics nationalism and internationalism relate to political or economic themes. As newspapers report by definition on current news, events such as strikes, riots, congresses, and meetings are fairly prominent. We can see a shift from an internationalism that before the First World War was dominated by conferences and world exhibitions to an internationalism that in the Interbellum was strongly linked to international agreements and conventions on labour and labour legislation. This can be related with the foundation of the International Labour Organization (ILO) in 1919 and overall institutionalisation of social internationalism. Another noticeable evolution that can be observed is that after the First World War associations related to sports, music, and cultural events (*wedstrijd, concert, voetbal, boxing, Kultuur, exposition*) become more prominent.

Fig. 6 Key associations with topic *internationalism* in *Vooruit* (1886-1938). Left: before the First World War. Right: 1918–38.

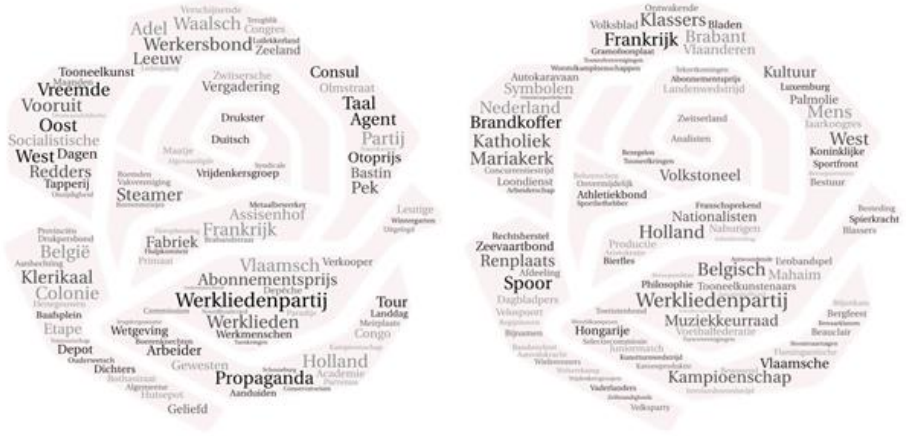

Fig. 7 Key associations with topic *nationalism* in *Vooruit* (1886–1938). Left: before the First World War. Right: 1918–38.

Fig. 8 Key associations with topic *internationalism* in *Le Peuple* (1886–1938). Left: before the First World War. Right: 1918–38.

Fig. 9 Key associations with topic *nationalism* in *Le Peuple* (1886–1938). Left: before the First World War. Right: 1918–38.

In both *Vooruit* and *Le Peuple* we see that references to countries associated with internationalism become more diverse over time, and refer to nations that are geographically more distant and globalized. Before the First World War, France, Germany, and the United Kingdom are often mentioned, whereas afterwards, words like *chineeschen, Slavische*, or *Roumanie, Colombia* occur. It is remarkable to see that in *Vooruit* references to colonialism in general, and Congo in particular, already occur significantly before the First World War. In *Le Peuple,* it is exactly the opposite.

Associations with the topic *nationalism* expanded over time and became more diverse. Early associations with nationalism expressed negative sentiments (such as *slavenland* in 1884), before and after the world wars there were strong expressions of patriotism (*vaderlandsliefde, gehoorzamen, opofferen*). A remarkable contrast between the associations with the topic nationalism and internationalism is that solidarity is only prominently linked to the latter. The same can be said for suffrage, although this word only occurs significantly in *Le Peuple*.

We also made a comparison between internationalism and cosmopolitanism. As we could expect, the rather liberal notion of cosmopolitanism (or variants such as *wereldburger*) barely occurs in *Vooruit*. And although the occurrence of cosmopolitanism

in *Le Peuple* is almost insignificant in terms of relative values, we found over 50 hits for some years without clear patterns of co-occurrence with other concepts.

Discussion and Conclusions

In this paper, we have used distant reading techniques to explore expressions of 'banal' socialist internationalism and national sentiments in two prominent Belgian socialist newspapers. The NER analysis showed a marked difference between the newspapers. The Ghent-based *Vooruit* showed a clear local focus, which is in line with the local neighbourhood sociability typical of Ghent socialism at that time. On the other hand, *Le Peuple* had a clearer international outlook, reflecting the Brussels socialists' cosmopolitan tradition. While both the Ghent and Brussels movements historically showed a strong affinity for the German workers' party, *Vooruit* pays more attention to France, the UK, and the Netherlands than Germany, whereas for *Le Peuple,* Germany was only preceded by France in frequency. The impact of the First World War (1917–18) is different for both newspapers. *Vooruit* loses its local focus to make more room for mentions of neighbouring countries as well as European countries, whereas *Le Peuple* gives more attention to neighbouring countries only, while retaining a strong focus on Belgium. This observation was confirmed by the collocation analysis, where there was a clear spike in nationalist topics in *Le Peuple* in 1917–18, while nationalism in *Vooruit* was less outspoken. Further research, ideally combining a mixture of distant and close reading techniques, is needed to confirm and explain these findings.

The collocation analysis further showed that the expressions of both banal nationalism and internationalism were mainly to be found in sport and cultural events. It is also striking that the associations with the progressive institutionalization of international socio-political life manifest themselves both in *Le Peuple* and in *Vooruit*. After the establishment of the International Labour Organization in 1919, this organization figured prominently within the internationalist topic, but not much more than the congress series and organizations established before the First World War. Notwithstanding the differences between the two periods, the First World War has not been a catalyst for a greater or lesser international orientation. Finally, there were not enough references to the concept of cosmopolitanism in the samples, either in a positive or in a negative sense, to be able to make statements.

Overall, the results are fairly consistent with what we already knew from previous research, but given the still experimental and time-consuming nature of these methods, such a confirmation is already meaningful. It means that these methods can also produce meaningful results for less studied periods. Of course, the explanatory value of the approaches described in this paper has its limitations. As has been argued before by many scholars, distant reading techniques must first and foremost be seen as a step that needs to be taken in order to know which text fragments need to be close read. This being said, it is important to stress the need to revise and refine the results of the text mining procedure at the lowest level of granularity, so that the historically relevant concepts the researcher is looking for throughout the entire corpus can be displayed in context, thus combining the power of distant and close reading.[26]

In addition to offering insights into the international scope of Belgian socialist newspapers, the work presented in this paper brought with it some methodological considerations. There would be no research without data, but even collecting said data

26 Stephen McGregor and Barbara McGillivray, 'A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell', *14th Conference on Natural Language Processing — KONVENS 2018* (Vienna, Austria: December 2018).

proved to be a laborious task. We believed selecting two-year samples of only two newspapers would make it more manageable, but that was hardly the case. Named-entity recognizers can, for example, take certain character or spelling differences into account, but if the OCR contains too many or unexpected changes, the names will not be recognized as such). Whilst this impact can be somewhat mitigated with bigger corpora, better OCR is needed for more fine-grained research.

We need powerful platforms for exploratory searching (Ngrams, word associations) and corpus building tools, such as the GREP-command that can be used to search and extract texts for lines or paragraphs containing a match with (a list of) words or search strings. This is one possible way to create a sub-corpus that allows a more fine-grained methodology. Proper corpus management indeed is the first, still difficult, hurdle to be overcome.[27] Future research could greatly benefit from ways to improve data acquisition and corpus exploration. This can be achieved by, on the one hand, fostering relationships with cultural heritage institutions to find better ways of cooperation that benefit all parties involved, and, on the other hand, building tools that make it easier for researchers to explore and evaluate their digital collections. Such tools should include ways to improve the OCR output, search through the collection, get an idea of words in context, and export interesting explorations and documents from the full collection. Intellectual workers of the world: avanti!

Joke Daems is a Postdoctoral Research Assistant Digital Humanities at Ghent University. She conducts research as a member of the Language and Translation Technology Team (LT^3) and offers support to the Ghent Centre of Digital Humanities (GhentCDH) on a faculty level, with a focus on digital language and text analysis.

Simon Hengchen is a Postdoctoral Researcher at the University of Helsinki and a lecturer at the University of Geneva. In his research, he focusses on computational methods to model lexical semantic change in noisy, multilingual data.

Tecle Zere holds a PhD in the Natural Sciences and a Postgraduate Diploma in ICT from the University of the Free State in South Africa. He has over 10 years experience, mainly as an ICT expert in the heritage sector, but also as a researcher and data expert. Tecle works for AMSAB Institute of Social History.

Bibliography

Baycroft, Timothy, *Culture, Identity and Nationalism: French Flanders in the Nineteenth and Twentieth Centuries* (Suffolk: Boydell & Brewer, 2004)

Billig, Michael, *Banal Nationalism* (London: Sage, 1995)

Bos, Maarten van den, and Hermione Giffard, 'Mining Public Discourse for Emerging Dutch Nationalism', *Digital Humanities Quarterly*, 10.3 (2016)

Brauer, Rene, and Matz Fridlund, 'Historicizing Topic Models, A Distant Reading of Topic Modeling Texts within Historical Studies', *International Conference on Cultural Research in the Context of 'Digital Humanities'* (St. Petersburg: Russian State Herzen University, 2013)

27 Maarten van den Bos and Hermione Giffard, 'The Grapevine: Measuring the Influence of Dutch Newspapers on Delpher', *Tijdschrift voor Tijdschriftstudies,* vol. 38 (2015), 29–41.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei, 'Reading Tea Leaves: How Humans Interpret Topic Models', *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (2009), pp. 288–96

De Bolla, Peter, *The Architecture of Concepts: The Historical Formation of Human Rights* (New York: Fordham University Press, 2013)

De Wilde, Max, and Simon Hengchen, 'Semantic Enrichment of a Multilingual Archive with Linked open Data', *Digital Humanities Quarterly,* 11.4 (2017)

Dogliani, Patrizia, 'European Municipalism in the First Half of the Twentieth Century: The Socialist Network', *Contemporary European History,* 11.4 (2002), 573–96

Edelstein, Dan 'Intellectual History and Digital Humanities', *Modern Intellectual History,* 13.1 (2016), 237–46

Gilleland, Michael, Gilleland, 'Levenshtein Distance, in Three Flavors', *Merriam Park Software* (2009)

Hill, Mark J., and Simon Hengchen, 'Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study', *Digital Scholarship in the Humanities*, in production, fqz024

Jockers, Matthew, and David Mimno, 'Significant Themes in 19th-Century Literature', *Poetics,* 41.6 (2012), 750–69

Jockers, Matthew, and Ted Underwood, 'Text-Mining the Humanities', in *A New Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens, and John Unsworth (Hoboken, NJ: Wiley-Blackwell, 2016), pp. 291–306

Koselleck, Reinhart, 'A Response to Comments on the Geschichtliche Grundbegriffe', *The Meaning of Historical Terms and Concepts* (1996), pp. 59–70

McGregor, Stephen, and Barbara McGillivray, 'A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell', *14th Conference on Natural Language Processing — KONVENS 2018* (Vienna, Austria:, December 2018)

Nelson, Robert, *Mining the Dispatch* [accessed 5 November 2018]

Newman, David, and Sharon Block, 'Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper', *Journal of the American Society for Information Science and Technology,* 57.6 (2006), 753–67

Rodriquez, Kepa, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska, 'Comparison of Named Entity Recognition Tools for Raw OCR Text', *Proceedings of KONVENS* (2012), pp. 410–14

Thompson, Paul, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou, 'Text Mining the History of Medicine', *PloS ONE,* 11.1 (2016), e0144717

Van Ginderachter, Maarten, 'Contesting National Symbols. Belgian Belle Époque Socialism Between Rejection and Appropriation', *Social History*, 34.1 (2009), 55–73

——, 'Social Democracy and National Identity: The Ethnic Rift in the Belgian Workers' Party (1885–1914)', *International Review of Social History,* 52.2 (2007), 215–40